DOCUMENT RESUME

ED 180 923 SO 012 304

AUTHOR Utano, Jack J.
TITLE Statistical Mapping by Computer.
PUB DATE Oct 79
NOTE 35p.: Paper presented at the Annual Meeting of the National Council for Geographic Education (Mexico City, Mexico, October 1979)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Computer Graphics; *Computer Programs; *Geography; Maps; *Statistical Data; Tables (Data)
IDENTIFIERS STATMAP Computer Program

ABSTRACT

The function of a statistical map is to provide readers with a visual impression of the data so that they may be able to identify any geographic characteristics of the displayed phenomena. The increasingly important role played by the computer in the production of statistical maps is manifested by the varied examples of computer maps in recent publications. The speed and versatility with which the computer is able to process locational information and generate corresponding graphic representations make it an invaluable mechanism for such wide ranging applications as experimenting with different ways of representing a geographic distribution, producing a map series for a regional atlas, or supplementing statistical analysis of spatial data. This paper discusses one such computer program named STATMAP (Statistical Mapping) which is able to generate the following statistical map representations on the CalComp vector plotter: 1) a continuous-tone choropleth map; 2) a graduated-circle map; 3) a graduated-bar map; and 4) a graduated-polygon map, where the latter three reflect special cases of the general graduated-symbol statistical map. There are three major sections to the report. In the first, the theoretical aspects concerning each of the four statistical maps are addressed. The second section presents a method for producing a "viewable" statistical map when skewed spatial distributions are invcived. An overview of the structural components of the computer program is provided in the last section. (Author/RM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED180923

FEB 11 1980

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

STATISTICAL MAPPING BY COMPUTER*

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Jack J. Utano
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Jack J. Utano
Instructor
Department of Geography
The University of Akron

SO 012 304

*Paper presented at the National Council for Geographic Education annual meeting in Mexico City, Mexico; October, 1979.

INTRODUCTION

Over the past few decades, the raison d'etre of statistical maps has been clearly enunciated in the cartographic literature. Dickinson, in his seminal work on the subject, postulated the following reasons regarding the utility of statistical maps (Dickinson, 1963: 12): "(1) to arouse greater interest in the subject matter presented; (2) to clarify it, simplify it or explain its more important aspects; (3) to prove a point referred to in text or speech; and (4) to act as a statistical 'quarry' for other users".

Augmenting more formal statements to these theoretical contentions, Jenks has argued that the function of a statistical map is to afford the reader with a graphic representation of statistical data in such a way that the spatial arrangements of the phenomena being displayed are manifested in a clear and unbiased manner (Jenks, 1963, 1976). It is only too well known that a scant minority of readers would be able to absorb correctly the spatial information contained in a censual table of enumerated data. A map provides a more effective communication medium for this type of information since relationships of distance and orientation or the spatial patterning of the phenomena are able to be distinguished more readily.

Moreover, the increasing important role of the computer in the production of statistical maps is equally manifested by the varied examples of computer maps in recent publications (e.g., see: Brassel and Utano, 1979a). The speed and versatility with which the computer is able to process locational information and generate corresponding graphic representations make it an invaluable mechanism for such wide-ranging applications as: experimenting with different ways of representing a geographic distribution; producing a map series for a regional atlas; or supplementing statistical analysis of spatial data (Brassel, 1977;

Brassel and Utano, 1979a; Brassel, Utano, and Hanson, 1977; Dudycha, 1978; Marble, 1979; Waugh and Taylor, 1976).

This paper focuses its discussion on one such computer program named STATMAP (STATistical MAPping) which is able to generate any of the following statistical map representations on the CalComp vector plotter: 1) a continuous-tone choropleth map; 2) a graduated-circle map; 3) a graduated-bar map; and 4) a graduated-polygon map, where the latter three reflect special cases of the general graduated-symbol statistical map. The report is partitioned into three major sections. In the first, the theoretical aspects concerning each of the four statistical maps are addressed. The second section presents a method for producing a "viewable" statistical map when skewed spatial distributions are involved. An overview of the structural components of the computer program is taken up in the last section.

THEORETICAL CONSIDERATIONS IN STATISTICAL MAPPING

The obvious utility of a statistical map is manifested by the information contained in table 1. Make an attempt to delineate the spatial information in the table. Most readers would experience considerable difficulty with such a task. Moreover, if the reader is provided with an "areal table", such as the one in figure 1, he would still have a difficult time in discerning the distributional characteristics of the data contained in tabular form. The function of a statistical map, then, is to afford the reader with a visual impression of the data so that he may be able to identify any geographic characteristics of the displayed phenomenon.

Although the advantages of representing statistical data (such as the data in table 1) in map form are numerous, one major problem in statistical mapping still persists. Once the potential map maker has decided to represent a given data distribution in map form, he must decide on the most appropriate statistical map representation to portray the distribution in such a way as to yield a faithful depiction of the data. This selection process has often become problematic for the wrong choice of graphic representation often yields a biased or inaccurate impression of the data distribution to the average map reader.

Thus, despite a seemingly universal consensus on the theoretical and practical purpose of statistical maps in general, a number of polemics have arisen within the cartographic discipline concerning two major facets in thematic map design: i) conventional symbol-subject association; and ii) information generalization. Although these two issues by no means represent mutually exclusive stages in the map design process, they have been treated separately in the literature due to their intractable theoretical components (for example, compare: Jenks, 1961; Tobler, 1973; Dobson, 1975; Chang, 1976). The major

STATE	%	STATE	%	STATE	%
ALABAMA	3.5	MAINE	0.7	OHIO	1.3
ARIZONA	44.2	MARYLAND	3.2	OKLAHOMA	3.6
ARKANSAS	9.9	MASSACHUSETTS	1.7	OREGON	52.5
CALIFORNIA	46.1	MICHIGAN	9.4	PENNSYLVANIA	2.4
COLORADO	35.6	MINNESOTA	6.6	RHODE ISLAND	1.2
CONNECTICUT	0.3	MISSISSIPPI	5.6	SOUTH CAROLINA	5.8
DELAWARE	3.2	MISSOURI	4.9	SOUTH DAKOTA	6.9
FLORIDA	10.4	MONTANA	29.6	TENNESSEE	6.7
GEORGIA	5.9	NEBRASKA	1.4	TEXAS	2.8
IDAHO	63.7	NEVADA	87.6	UTAH	65.1
ILLINOIS	1.6	NEW HAMPSHIRE	12.3	VERMONT	4.8
INDIANA	2.1	NEW JERSEY	2.7	VIRGINIA	9.2
IOWA	0.6	NEW MEXICO	33.5	WASHINGTON	29.1
KANSAS	1.4	NEW YORK	0.8	WEST VIRGINIA	7.8
KENTUCKY	5.4	NORTH CAROLINA	6.5	WISCONSIN	5.3
LOUISANA	3.7	NORTH DAKOTA	5.3	WYOMING	48.6

TABLE 1: Percentage of individual state's total land acreage that is federally owned as of 1977.

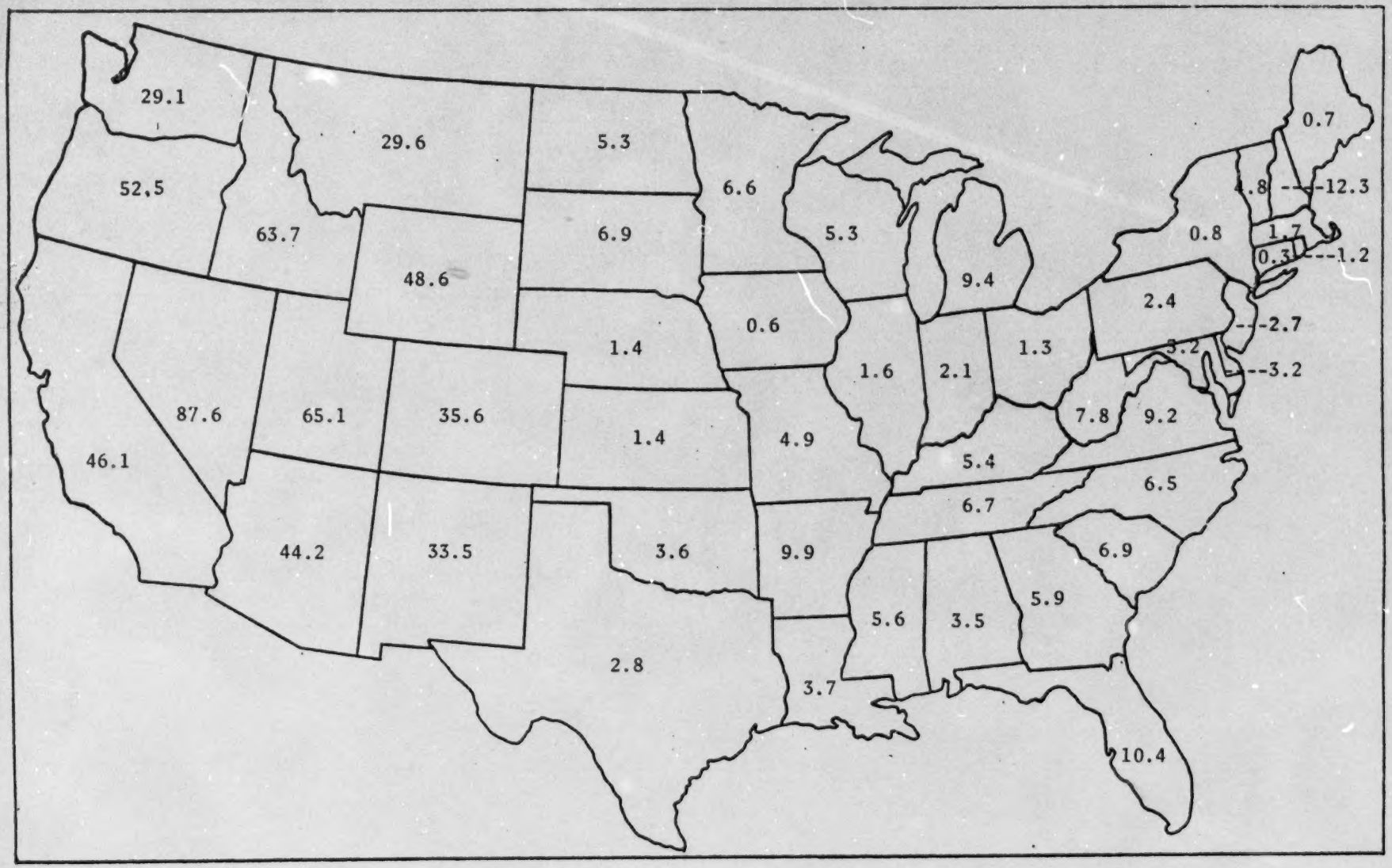


FIGURE 1: Percentage of individual state's total land acreage that is federally owned as of 1977.

problems of both are taken up in the following sections.

Conventional Symbol-Subject Association

Statistical data bearing either an implicit or explicit spatial component are normally partitioned into four types: point, line, area, and volume (Wright, 1955). Distinction of these spatial data types depends upon the map scale and level of generalization chosen and are used according to how the statistical data are to be represented on a map. For example, if cities are to be displayed as points, city populations may then be considered point data; if cities are to be represented as having areal extent, population densities in the areal unit subdivisions within the city are considered area data (Muehrcke, 1972). The following discussion focuses upon the more common point and area data types together with their associated symbol-subject conventions.

Generally, there are four symbol forms to display two-dimensional spatial information in which the use of these symbols is governed by the specific function assigned in thematic map convention (Raisz, 1962; Robinson and Sale, 1969). Point symbol maps are used to depict point data that contain discrete locational aspects. A unit value is assigned to some point, usually a dot, and a proportionate number of point symbols are positioned within the approximate geograhic location of the subject matter (Dahlberg, 1967). A typical illustration of this kind of point symbol representation is a dot map portraying the geographic frequency of various populations, e.g. sheep farms, fruit stands.

When point data exist in geographic clusters and are large in number, proportionate or graduated-symbols, usually circles (Flannery, 1971), are chosen to represent an aggregation of the point distribution. A point symbol is placed at the approximate center of the cluster of points and the size of the symbol is drawn as a function of the magnitude of the point aggregation at

the location. Town and city population or manufacturing output, for example, are amenable to graduated-symbol representation. Given the population of a city, a symbol size is derived with respect to the city's position in the distribution of city populations being displayed. In practice, graduated-symbols have also been used to represent aggregate point data that refer to areal statistical units (Chang, 1976). In these circumstances, each unit is seen as "only a locational or point value for mapping purposes" (Robinson and Sale, 1969: 124-5). In this way, proportionate symbols may be employed to serve as area symbols to map population or manufacturing output related to various enumeration districts, e.g. minor civil divisions, political wards, census tracts.

Statistical data that exhibit properties of areal extent are commonly represented by area symbols, usually of the choropleth variety, in which the attribute being considered is assumed to be evenly dispersed over the areal units. In choropleth maps, monochromatic patterns are used to symbolize the magnitudinal distribution of the statistical data in such a way that a gray tone density (percent of area inked) is assigned according to the datum value of the respective areal unit (Willams, 1958; Jenks and Knos, 1961). Given the assumption of equal distribution over area, choropleth maps require an areal standardization of the statistical data, usually expressed in percentage or ratio form, so that the effect of area size is eliminated.

Lastly, statistical data in which the distribution may be conceived as ubiquitous or continuous but varying in intensity are normally represented by isarithmic area symbols. In this context, isarithms are defined as "lines which connect places of equal value" (Raisz, 1962: p 199). A contour map, for example, are lines linking places of equal elevation. A wide array of statistical data, distributions like air temperature, pressure, elevation, etc., are really three-

dimensional in nature and may also be represented by volume symbols. This type of data is subsumed under the concept of 'statistical surface' (Robinson, 1961; Schmid and MacCannell, 1955). However, when the preservation of the planimetry of the areal distribution is desired, two-dimensional representations, i.e. area symbols, is the preferred mapping technique (Jenks, 1961; Robinson and Sale, 1969: 152-164).

Despite these generally accepted symbol-subject conventions, Dobson has questioned the validity of establishing thematic map conventions in providing certain symbol types to represent specific kinds of statistical data (Dobson, 1975: 53):

> Map readers, unfortunately, are non-cartographers (indeed sometimes non-geographers) who have only a rudimentary map reading background and who are, therefore, unfamiliar with the symbol conventions the cartographer uses. The essence of this dichotomy revolves around thematic maps as a communicative device and especially the error that may be involved due to conceptual differences between the map author and the map reader relating to the symbolization used.

Dobson goes on to claim that if the map reader does not associate the specific symbol convention given a particular type of statistical distribution or his symbol-subject association does not correspond with accepted thematic practice, the communicative effectiveness of the statistical map is reduced substantially if not forfeited. He buttressed his arguments with a study examining the relationship between map symbol convention and map reader experience. The experiment involved testing 'trained' and 'untrained' map readers. The results acknowledged that the trained reader can think symbol-subject matter relationship as well as recognize them in map format. Conversely, the inexperienced reader has only limited success in either of these ventures. Dobson concluded that since the untrained reader generally experiences severe difficulty in 'decoding'

correctly the conventional symbol-subject relationship "cartographers should either attempt to devise new standards or more effectively to disseminate the method of interpretation those now employed" (Dobson, 1975: 63). Jenks addressed this issue in his statement that despite the advantages that might be gleaned from the development of new informational concepts and types of statistical maps, the cartographer should adhere to the recommended thematic map conventions (Jenks, 1976: 14); He continues:

> Each mappable phenomenon has distinctive geographical characteristics and an understanding of these is the basis upon which symbolic conventions were established. It is also clear that many map users expect certain types of phenomenon to be depicted in a certain manner, and if the symbolic language of the map coincides with the expectations of the reader the impact of the map message occurs more rapidly and with fewer aberations than would the case otherwise.

Balchin, in his introduction of the term 'graphicacy', lends strong support to Dobson's contention that cartographers need to better educate society in the science of map reading (Balchin and Coleman, 1965; Balchin, 1976). Indeed, Balchin stresses the need, with an analogy to literacy, for teaching the fundamentals of map reading before mapmaking (Balchin, 1976: 36).

Aside from the choice of symbol representation, however, the cartographer is faced with two additional and perhaps equally important components in statistical map design. Withstanding the problem of data classification which is taken up in the next section, the level of measurement of the data (nominal, ordinal, interval, or ratio) when cross-classified with spatial data type (point, line, area, or volume) influences the selection of not only the symbol representation but also the symbol characterization, namely, size, shape, color, or pattern (Muehrcke, 1972: 14-16; Robinson and Sale, 1969: 95-101). A choropleth map, for example, depicting areal land use types would give the reader a false impression of explicit magnitude in a distribution that is really

nominal or at most ordinal in scale. The communicative effectiveness, i.e. the minimization of map reader error, of the map is thus predicated on both the cartographer's choice of symbol and the manner in which the symbol is characterized.

The issue of symbol-subject convention is therefore subsumed in the broader question of semiotics, the theory of signs, vis-a-vis thematic map standardization (compare Bertin, 1967; Arnberger, 1974; Robinson, 1974). The study of semiotics involves three major aspects: relationships of map symbols with subject matter (called syntactics); relationships of map symbols and their denotation (semantics); and the relationships of map symbols and the map user (pragmatics). Unfortunately, semantics and pragmatics have been given scant attention (compare Dacey, 1970; Morrison, 1974). The issue of symbol-subject convention will remain unresolved until proper attention is devoted to these latter semiotics aspects.

Despite this problem, the issue of symbolic conventions appears to be overstated for two reasons. First, any rigid form of symbol standardization must be challenged by the contention that it would necessarily lead to the inhibition of new or unanticipated kinds of map expression and foster the discouragement of experimenting with alternative symbolisms. As Muehrcke has pointed out: "map symbolism is perfectly general and should be designed or selected to suit the purposes of cartographic analysis" (Muehrcke, 1972: 38).

Secondly, regardless of what symbol-subject relationship employed, the ability of the statistical map to serve as a viable medium for transmitting information from the cartographer to the map reader rests heavily on a logical and well thought out legend. Since the legend identifies the symbol elements and their attendant relational and numerical associations on the map, the cartographer can greatly enhance or negate the communicative purpose of

the statistical map by the way the symbol components are delineated in the legend. Although the size of the legend is obviously subordinate to that of the map body, its importance in linking the cartographer's message that is manifested in the statistical map to the reader cannot and should not be usurped. Just as an abstract is intended to outline the basic components and theme of the article, the legend is responsible for identifying the basic symbolic components and message of the statistical map. Unfortunately, as with most abstracts, legends are probably constructed as an afterthought and in their hasty construction, the interfacing function of the legend is oftentimes compromised. However, if the legend provides the user with a complete documentation of the symbol elements and their relationships displayed on the map, the issue of symbol-subject convention appears to be vacuous.

Information Generalization

A recurrent theme threaded throughout the literature on statistical mapping is the insistence that statistical maps are but symbolized generalizations of the information contained in its tabular counterpart (Brassel and Utano, 1979b; Peterson, 1979). Based on the premsis that simplicity reflects a fundamental dictum in cartography, many cartographers have argued that the information provided by a data array cannot and should not be mapped on a one-to-one correspondence; the data transformation process from table to map necessarily involves a classification stage of the N values in the distribution into a much smaller number of groups. Such a classification inevitably leads to some sort of information loss (compare Jenks, 1961, 1963; Robinson and Sale, 1969: 164-170; Monmonier, 1976: 25-34). Jenks concurs with his statement that "statistical maps tend, therefore, to be inefficient and inaccurate sources of data" (Jenks, 1976: 12). Despite the information loss attributed to statistical maps, their

utility in portraying the spatial information are not negated; consider Jenks and Coulson (Jenks and Coulson, 1963: 119):

> If we assume a degree of scientific integrity in these cartographic endeavors, the intent of the map-maker is to present a 'reasonably' useful representation of the distribution.... If these intervals are wisely chosen the reader may gain a clearer understanding of the areal relationships than he could have obtained from the original data, but when the categories are poorly selected he may gain only distorted or inaccurate impressions of the distribution. The problem of selecting classes is, therefore, the most important phase of constructing a statistical map since in this step the map-maker controls map interpretation.

Choropleth maps, in particular, have been based traditionally on the assumption that the reader should be provided with a set of class intervals that yields an optimal reflection of the natural breaks present in the data (Monmonier, 1975). Given a data array, the problem at hand is to determine both the optimal number of classes and the range of the class intervals to represent the distribution in an unbiased manner. Work on these points has been extensive to say the least (compare, for example: Mackay, 1955; Scripter, 1970; Jenks and Caspall, 1971; Monmonier, 1972, 1973, 1975, 1976; Chang, 1974; Dobson, 1974; Muller, 1976; Evans, 1977; Wasilenko and Moelleving, 1977; Peterson, 1978, 1979).

Unfortunately, there exists a proclivity of 'feasable solutions' in determining distribution breaks for class intervals and, to date, no universal 'correct' method has been accepted (Monmonier, 1976). One only has to examine the contents of recent issues in any of the major cartographic journals to discover that someone has implemented yet another algorithm for 'optimizing' class intervals for statistical maps, particularly of the choropleth variety.

Tobler has brought forth an interesting solution to the class interval problem (Tobler, 1973). He has developed a computer program that can generate

gray tones anywhere along a white to black continuum (Tobler, 1971). Given N data values, the program constructs N shading densities, thus emancipating the cartographer from the troublesome burden of establishing class intervals and eliminating what Tobler calls the 'quantization error' of discrete-symbolized choropleth maps. Dobson has retorted, however, that increasing the amount of information on the map in this manner must of necessity decrease the ability of the reader to comprehend it. Whereas the quantization error declines as the number of classes approaches N, the perceptual error increases (Dobson, 1973; Jenks, 1976).

There appears to be two separate but highly correlated aspects to this issue: i) information loss and error; and ii) map reader loss and error. Although information generalization necessitates information loss, many cartographers would agree that data classification is essential for rendering a statistical map 'useful' to the map reader. Indeed, many cartographers would stress that statistical maps are designed and intended to portray a general picture of the nature of the spatial distribution and that they are not suitable media for precision work (Muehrcke, 1972: 38). One practical shortcoming of this reasoning is that a number of arithmetically and computationally accurate maps depicting the same data set can and most often will produce different distributional patterns. It is therefore quite possible and highly probable that a statistical map formulated within some class interval framework would lead to a deliberate graphic falsehood thereby causing the map user to misinterpret the true spatial distribution of the phenomenon. The critical question then becomes: which classification scheme should the cartographer use that best exemplifies the data.

Another form of induced map reader error may arise from the user getting the impression that the areal units grouped in a given class interval represent

homogeneous or similar data values within the distribution (Bachi, 1974). Classification of a skewed distribution testifies to the inadequacy of this assumption. Despite a class interval method which would maximize inter-group variance and minimize intra-group variance, observations along the skewed portion of the distribution curve would likely be lumped into the same class interval. In the case of a choropleth display, the reader would perceive these areal units depicted by the same shading density as being 'equal' when in fact the observation values may be quite disparate.

Tobler's elimination of information generalization in a choropleth map renders the class interval problem academic. It has been criticized, however, on the grounds that this type of two-dimensional continuously-shaded statistical surface renders the map unreadable. This criticism seems to stem from the sample map provided by Tobler (Tobler, 1973: 263). The problem, however, lies not with the map body but with the legend. Even a 'trained' map reader would experience difficulty in identifying the relational and numerical associations on the map given the weak structure of Tobler's legend. Brassel and Utano have developed a computer program that incorporates Tobler's shading density algorithm but creates a legend that greatly enhances the map user's readability (Brassel and Utano, 1978a & 1979a). The legend contains a grey tone scale that is 'mapped' to a discrete set of evenly-spaced density symbols together with their respective numeric values. In this way, the reader is able to retrieve from the map accurate numerical relations in addition to generalized visual regionalizations.

As emphasized earlier, internal documentation of a statistical map holds the key to the success of the communicative effectiveness of the map. Both issues of symbol-subject conventions and information generalization are circumvented if a well-structured legend is attached to the map body. Thus, the critical issue that needs to be tackled when representing statistical data in map form

is the integration of the map body with various annotative features to enhance the communicative ability of the statistical map.

Program STATMAP

Program STATMAP has been designed specifically to address this problem in statistical mapping. Given a set of locational data describing the areal boundaries and a set of statistical data denoting the values to be represented, the program generates a CalComp plotter realization of any of the following statistical maps:

(1) Continuous-Tone Choropleth Map: The areal units are shaded proportional to the range of the data value distribution and the shading density distribution. The data value distribution is supplied by the user in the form of a data file. The shading density distribution is also determined by the user who assigns a maximum shading density (% area-inked) to the maximum data value and a corresponding minimum shading density to the minimum data value.

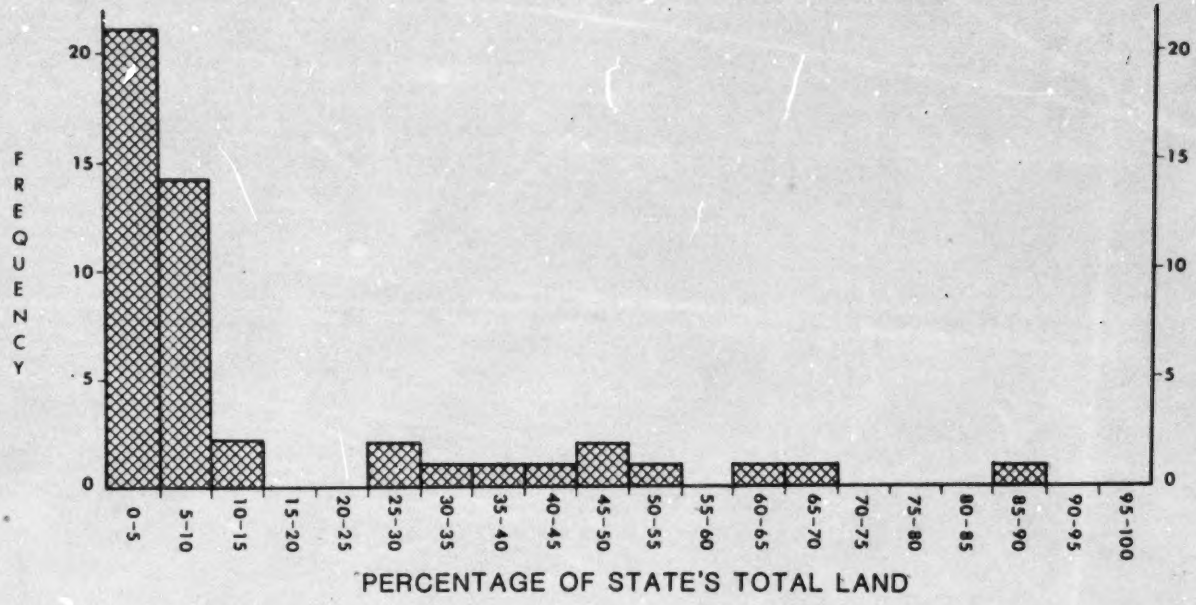
(2) Graduated-Circle Map: At the center of each areal unit, a circle is drawn whose size is established by the value designated for the unit with respect to the data value range in the distribution. The circle may be completely open or be filled with a number of progressively smaller concentric circles. Maximum radius size and the distance between the concentric circles are user options. Circle overlap, however, is not corrected in this routine.

(3) Graduated-Bar Map: At the center of each areal unit, a rectangular bar is drawn whose size is determined by the value associated with the unit with respect to the data value range in the distribution. Maximum bar height and the shading density of the graduated-bar are supplied by the user. Bar overlap, however, is not corrected in this routine.

(4) Graduated-Polygon Map: At the center of each areal unit, a polygon is drawn whose shape is similar to the original unit, and whose size is determined by the value designated for the unit with respect to the data value range in the distribution. The shading density of the graduated-polygons is furnished by the user. Polygon overlap, however, is not corrected in this routine. This type of statistical map is an extended version of the noncontiguous cartograms proposed by Olson (1976).

Symbolism Determination

Symbolism selection is constructed in a continuous manner. For the choropleth map, the data values pertaining to the areal units are assigned a shading symbolism according to the following scheme. The minimum and maximum data values in the distribution are determined and the zone, or zones, with the highest value are shaded with the darkest symbolism (usually 100% of the area is covered--i.e., BLACK). Likewise, the zone, or zones, with the lowest data value are shaded with the lightest symbolism (usually 0% of the area is covered--i.e., WHITE). Intermediate values are then designated a shading density proportionally between the BLACK-WHITE continuum according to a non-linear psychophysical power-function (Tobler, 1973; Brassel and Utano, 1979b; Peterson, 1979). This symbolism scheme is similar for the graduated-symbol maps. The extreme data values in the distribution are found and the data zones pertaining to the extremes are represented by the largest and smallest graduated-figurative symbols respectively (circle, bar, or polygon). Intermediate values are then scaled accordingly.


FREQUENCY
20
15
10
5
0
0-5
5-10
10-15
15-20
20-25
25-30
30-35
35-40
40-45
45-50
50-55
55-60
60-65
65-70
70-75
75-80
80-85
85-90
90-95
95-100
PERCENTAGE OF STATE'S TOTAL LAND

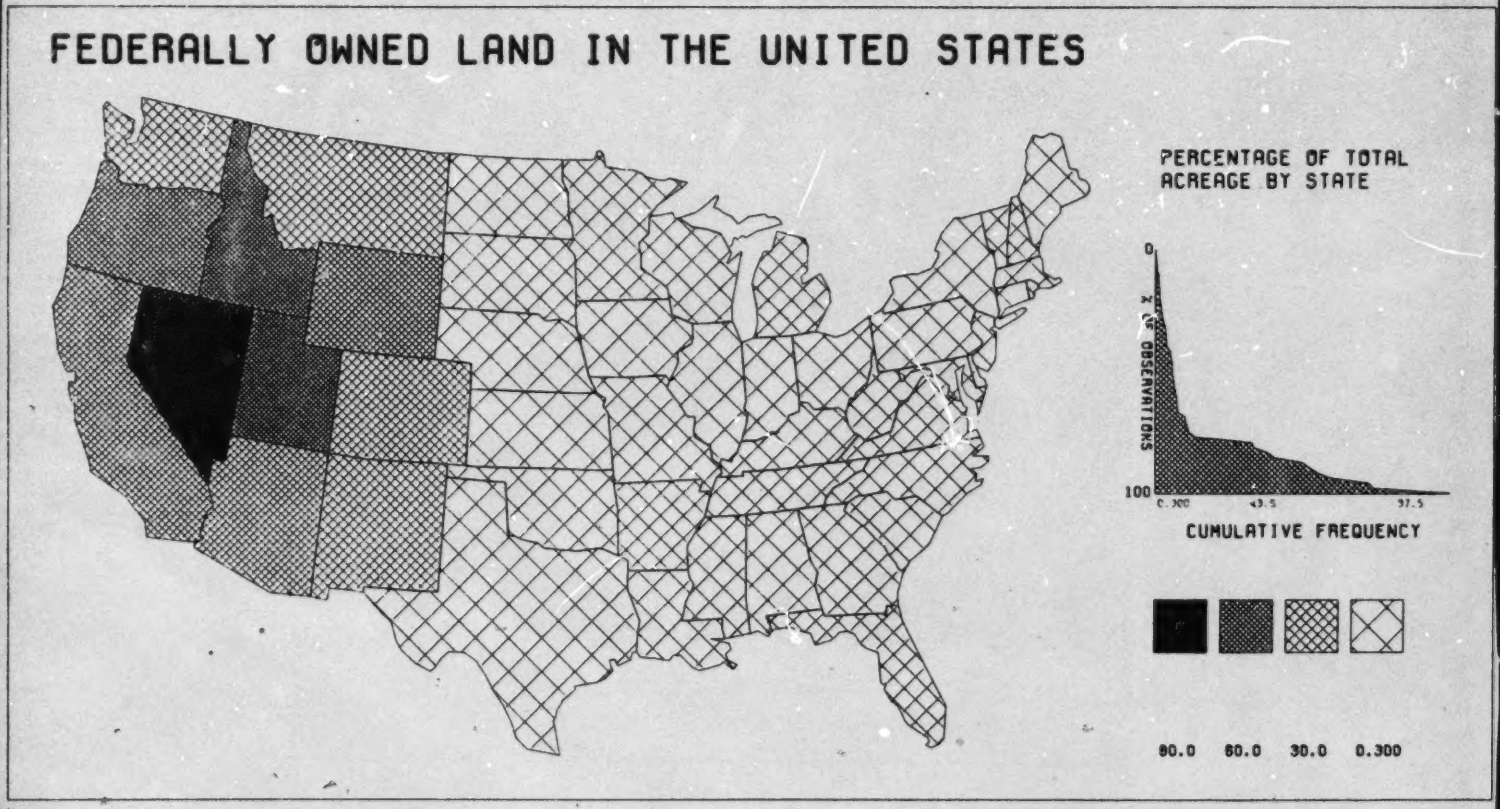

FEDERALLY OWNED LAND IN THE UNITED STATES
PERCENTAGE OF TOTAL ACREAGE BY STATE
0
100
CUMULATIVE FREQUENCY
80.0
60.0
30.0
0.300

Figure 3

REPRESENTING SKEWED SPATIAL DISTRIBUTIONS

In applications where the shape of the statistical data frequency distribution does not conform to a conventional normality arrangement, the continuous symbolism assignment scheme becomes problematic. Consider figure 2 which shows a histogram of the data distribution depicted in table 1. The graph hardly resembles a normal frequency distribution. 35 of the 48 or nearly 73% of the data values lie within the first 10% of the frequency distribution. The mean is , whereas the median equals , thereby indicating a sharp negative skew in the spatial distribution. The problem of symbolizing this particular distribution in a continuous manner becomes very clear when we map the distribution. Figure 3 represents the graphical product when the distribution is represented within a strict continuous symbolism procedure. As can be clearly gleaned from the map, the large absolute deviation between the extreme data values relative to the remaining distribution tends to mask or at least underscore the symbolism range of the residual distribution. To compensate for this problem a quasi-continuous allocation has been implemented in STATMAP to enhance the visual communication of the four statistical maps when plotting skewed data sets (Brassel and Utano, 1978a, 1979b).

Quasi-Continuous Symbolism Determination

The quasi-continuous algorithm is operationalized in the following manner. Two additional data parameters are required from the user to invoke this option -- one value that is less than the maximum data value (denoted as the maximum cutoff value) and one that is greater than the minimum data value in the statistical distribution (minimum cutoff value). In the case of a choropleth map, those areal units having data values greater than the maximum cutoff value are allocated the highest shading density (BLACK). Conversely, those areal units having data

values less than the minimum cutoff value are symbolized with the lowest shading density (WHITE). The continuous symbolism allocation is then applied to the data values between the cutoff extremes. In the case of the graduated-symbol maps, the quasi-continuous scheme permits the user to allocate maximum and minimum height-symbols to extreme values and use the continuous algorithm to the residual distribution. The program provides an option to generate a lineprinter graph of the data value distribution to help the user determine these cutoff values.

Figure 4 provides a schematic illustration of this symbolism scheme. The program parameters for STATMAP would be:

ZMIN, ZMAX -- the minimum and maximum data values in the data distribution, or the external class limits of the data space;

ZBOT, ZTOP -- the minimum and maximum cutoff values in the data distribution, or the internal class limits of the data space;

WHITE, BLACK -- the respective shading densities or percent ink coverage for areal units having data values between the internal and external class limits, i.e. between ZMIN-ZBOT and ZTOP-ZMAX;

LIGHT, DARK -- the shading density limits for the data values between the internal class limits, i.e. between ZBOT-ZTOP.

This quasi-continuous scheme represents a compromise solution between Tobler's (1973) continuous technique and the traditional discrete classification approach. The three map symbolism schemes are compared graphically in Figure 5 and an example map based on this quasi-continuous method is shown in Figure 6.

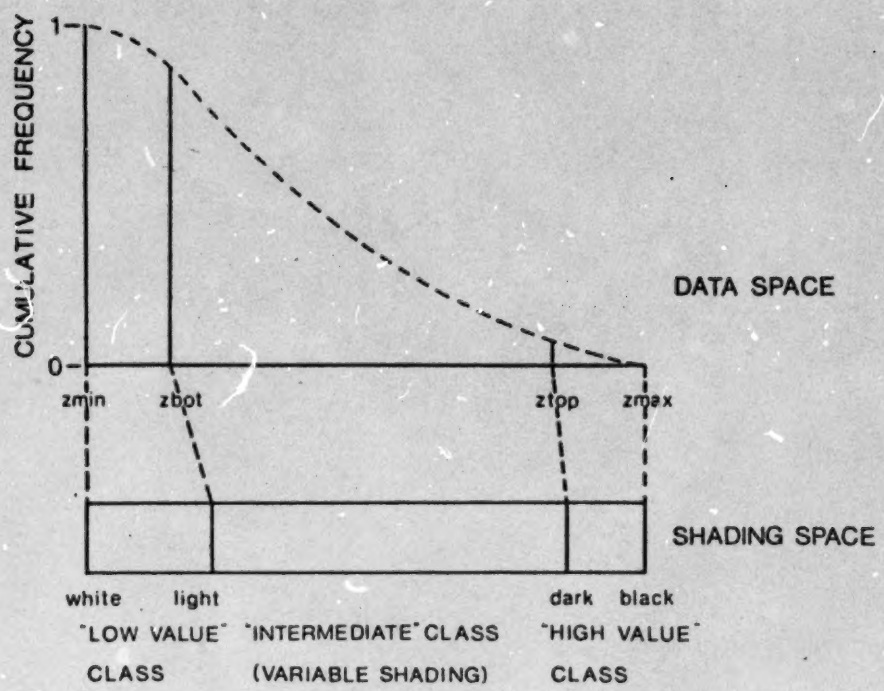


Figure 4

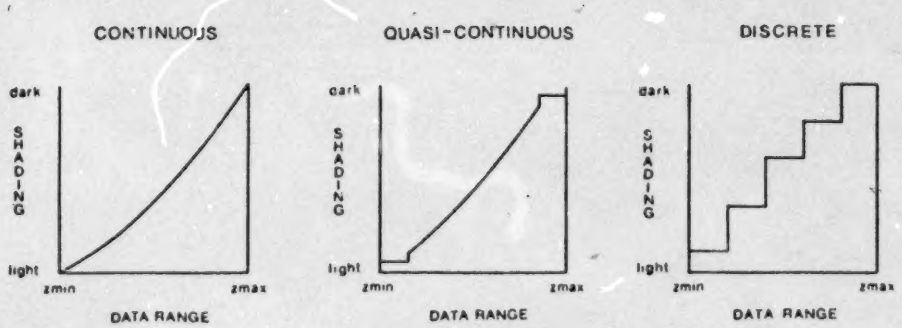


Figure 5

A STRUCTURAL OVERVIEW OF STATMAP

The primary processing steps of the program are given in Figure 7. For the production of any of the four statistical maps, STATMAP requires three types of user input:

(1) A map type and a set of plotting options (user-supplied parameters)

(2) the data values of the spatial distribution (statistical data)

(3) the area base map in digital form, i.e. X-Y coordinates of the areal unit boundaries (locational data)

Once the map type specification is designated (this procedure is described in more detail in the next section), two kinds of user-supplied parameters are required:

(A) A set of global parameters is necessary to establish options common to all four statistical maps. Variables such as the plot height, plot width, plot scale factor, pen color and pen width are examples of the global parameter section.

(B) A set of local parameters is needed to generate a particular statistical map. Some examples include: the highest and lowest shading density for the construction of a continuous-tone choropleth map; maximum and minimum radii for the graduated-circle map; maximum bar height and width for the graduated-bar map; and, the polygon shading density for the graduated-polygon map.

In addition to the user-supplied parameters, two types of data are required by STATMAP:

(A) The user must supply a file of 'locational data' that describes the X-Y coordinates of the areal units to be represented. For example, X-Y coordinates,

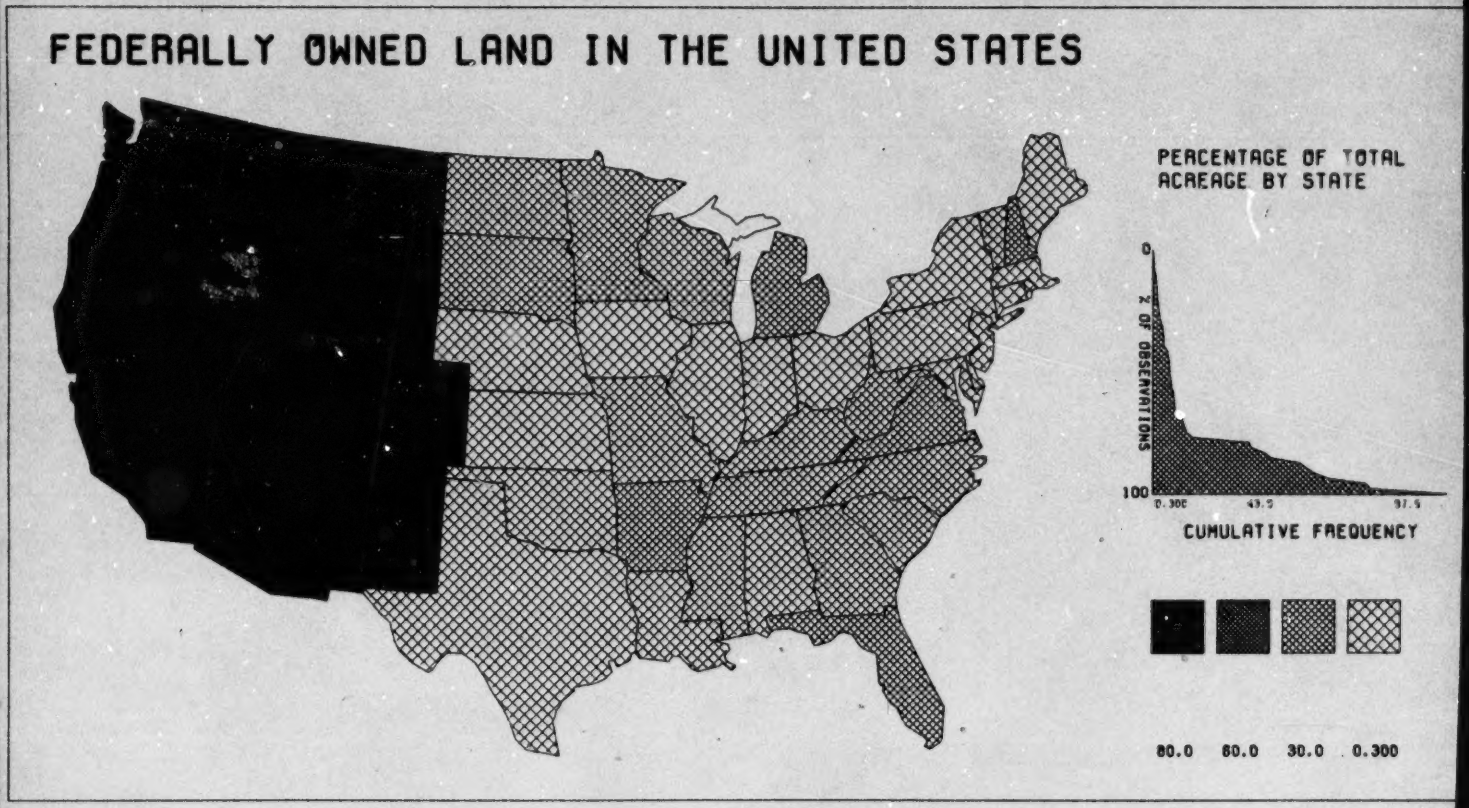

FEDERALLY OWNED LAND IN THE UNITED STATES
PERCENTAGE OF TOTAL ACREAGE BY STATE
0
% OF OBSERVATIONS
100
0.300
49.5
97.5
CUMULATIVE FREQUENCY
80.0
60.0
30.0
0.300

Figure 6

STATMAP : Program Processing Steps.

— READ MAP TYPE SPECIFICATION RECORD

Choropleth — Graduated - Symbol

— READ USER-SUPPLIED PARAMETERS

Global — Local

— CHECK USER - SUPPLIED PARAMETERS

Value Range Verification

— READ DATA FILES

Statistical — Locational — Centroids (optional)

— COMPUTE MAPPING PARAMETERS

Extreme Data Values
Extreme X-Y Coordinates
Plot Scale Factor

— ASSIGN SYMBOLISM TO THE AREAL UNITS

Continuous — Quasi - Continuous

— PLOT MAP BODY

Areal Unit Outlines — Symbols

— PLOT ANNOTATIVE FEATURES

Titles — Legend — Frequency Curve

in any grid referencing system, of the United States depicting state county or census tract boundaries may serve as appropriate locational data.

(B) The user must supply a file of 'statistical data' that describes the data value of each of the areal units; e.g. average income by U.S. states, counties, or city census tracts. Given N number of areal units defined in the locational data file, there must be N number of data values in the statistical data file. In addition, there must be a one-to-one correspondence between the sequential order of the locational data and statistical data. In other words, the first data value in the statistical data file must pertain to the first areal unit set of X-Y coordinates in the locational data file; e.g. if the first data value is given for average income in Alabama, then the locational data file must contain the X-Y coordinates for Alabama first, etc.

The program also permits additional non-valued polygons to be plotted on the map body. This option is useful for showing additional spatial information while not using these non-valued polygons in the statistical compilation of the map. Plotter output from STATMAP includes a number of annotative features designed to enhance map readability. A legend is plotted with symbolism representative of various discrete slices in the statistical distribution. A special rounding algorithm has been designed to find the most even data values and to associate the discrete legend boxes or symbols with the appropriate rounded values. A cumulative frequency curve and an improved lettering procedure (Brassel and Utano, 1978b) are also included. Examples of each of the four statistical maps are provided in Figures 8-11.

Modular Processing Steps

- STATMAP was written in FORTRAN IV and most of the language code is in

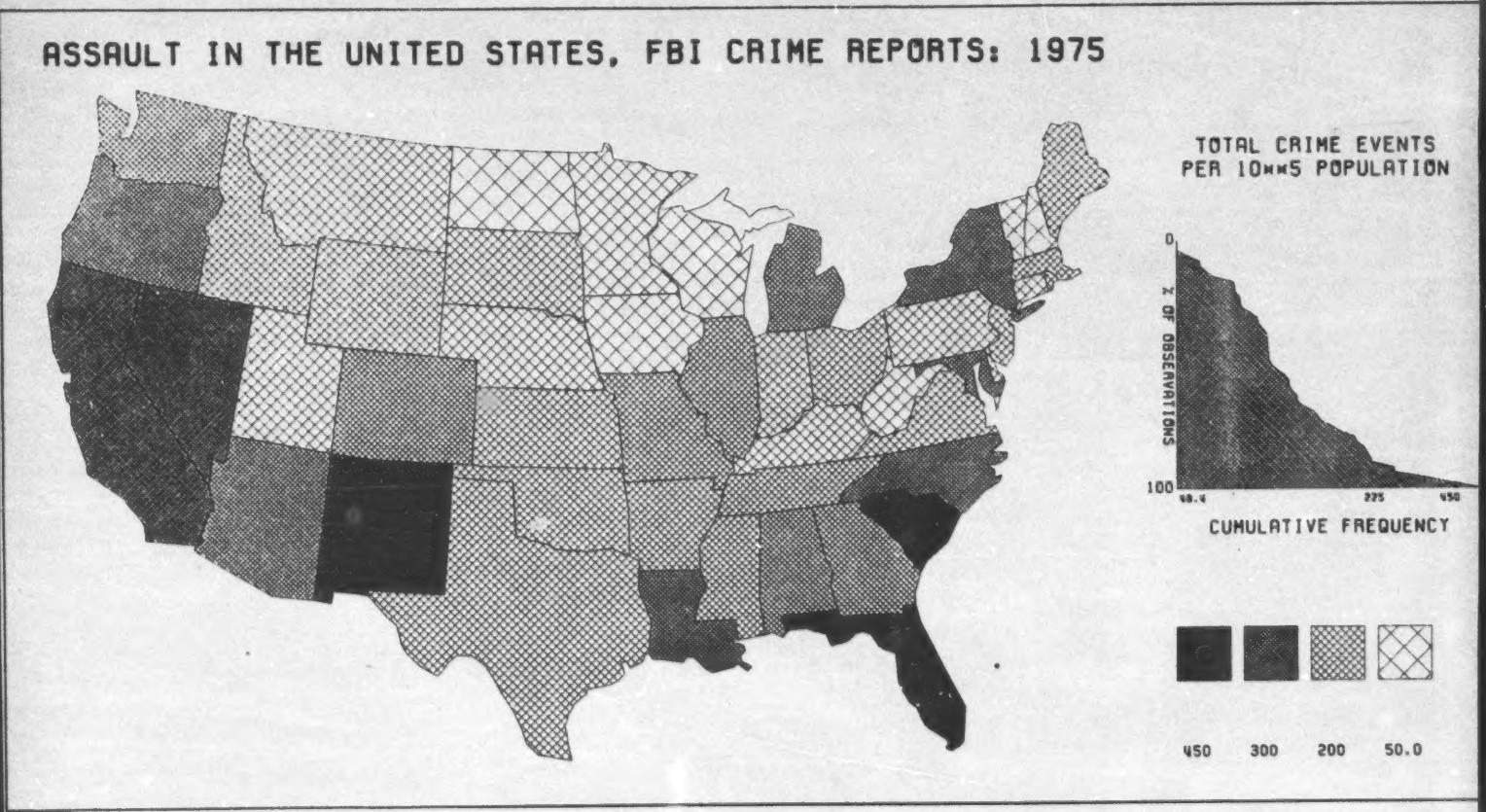

ASSAULT IN THE UNITED STATES, FBI CRIME REPORTS: 1975
TOTAL CRIME EVENTS
PER 10##5 POPULATION
0
% OF OBSERVATIONS
100
225
450
CUMULATIVE FREQUENCY
450
300
200
50.0

Figure 8

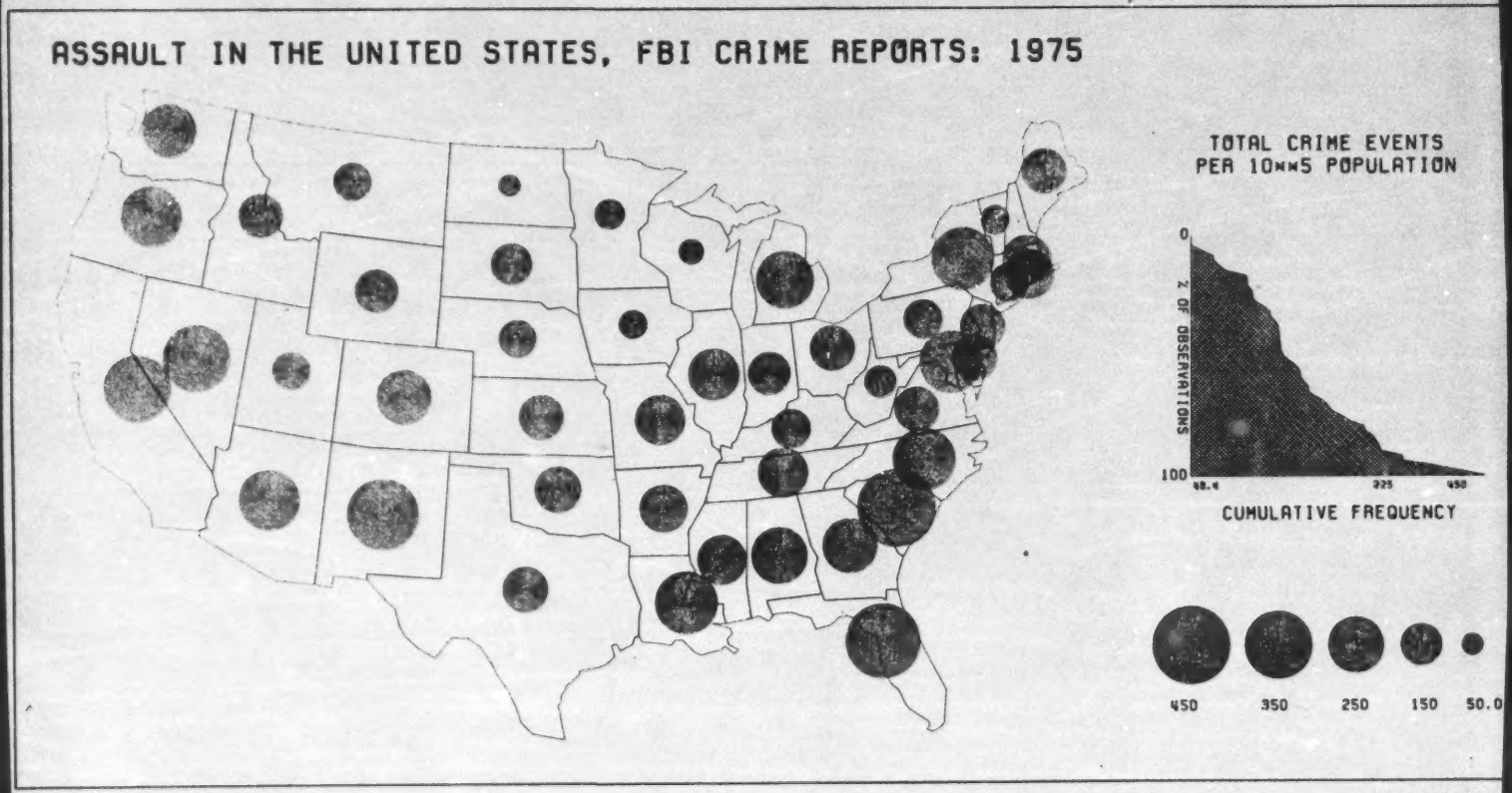

ASSAULT IN THE UNITED STATES, FBI CRIME REPORTS: 1975
TOTAL CRIME EVENTS
PER 10##5 POPULATION
0
% OF OBSERVATIONS
100
225
450
CUMULATIVE FREQUENCY
450
350
250
150
50.0

Figure 9

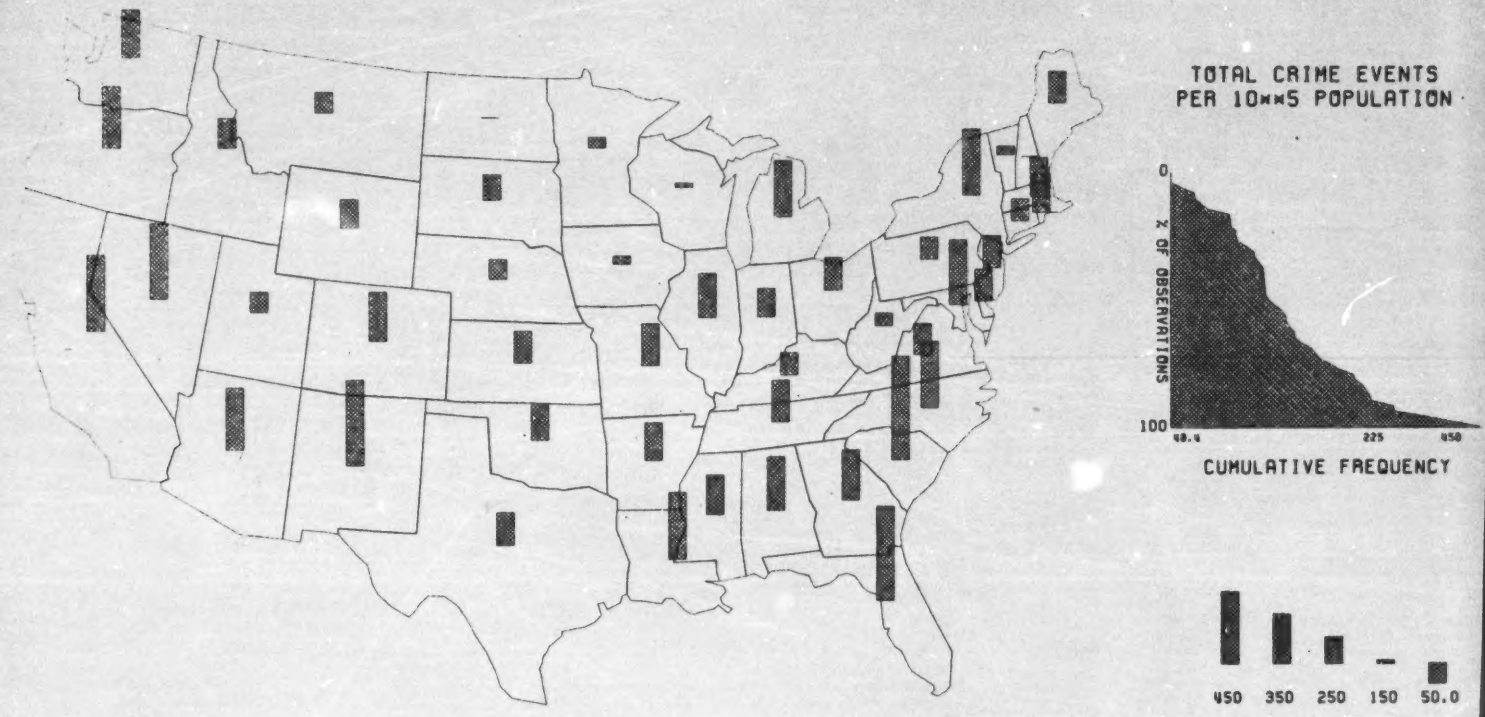

ASSAULT IN THE UNITED STATES, FBI CRIME REPORTS: 1975
TOTAL CRIME EVENTS
PER 10**5 POPULATION
0
% OF OBSERVATIONS
100
48.4
225
450
CUMULATIVE FREQUENCY
450
350
250
150
50.0

Figure 10

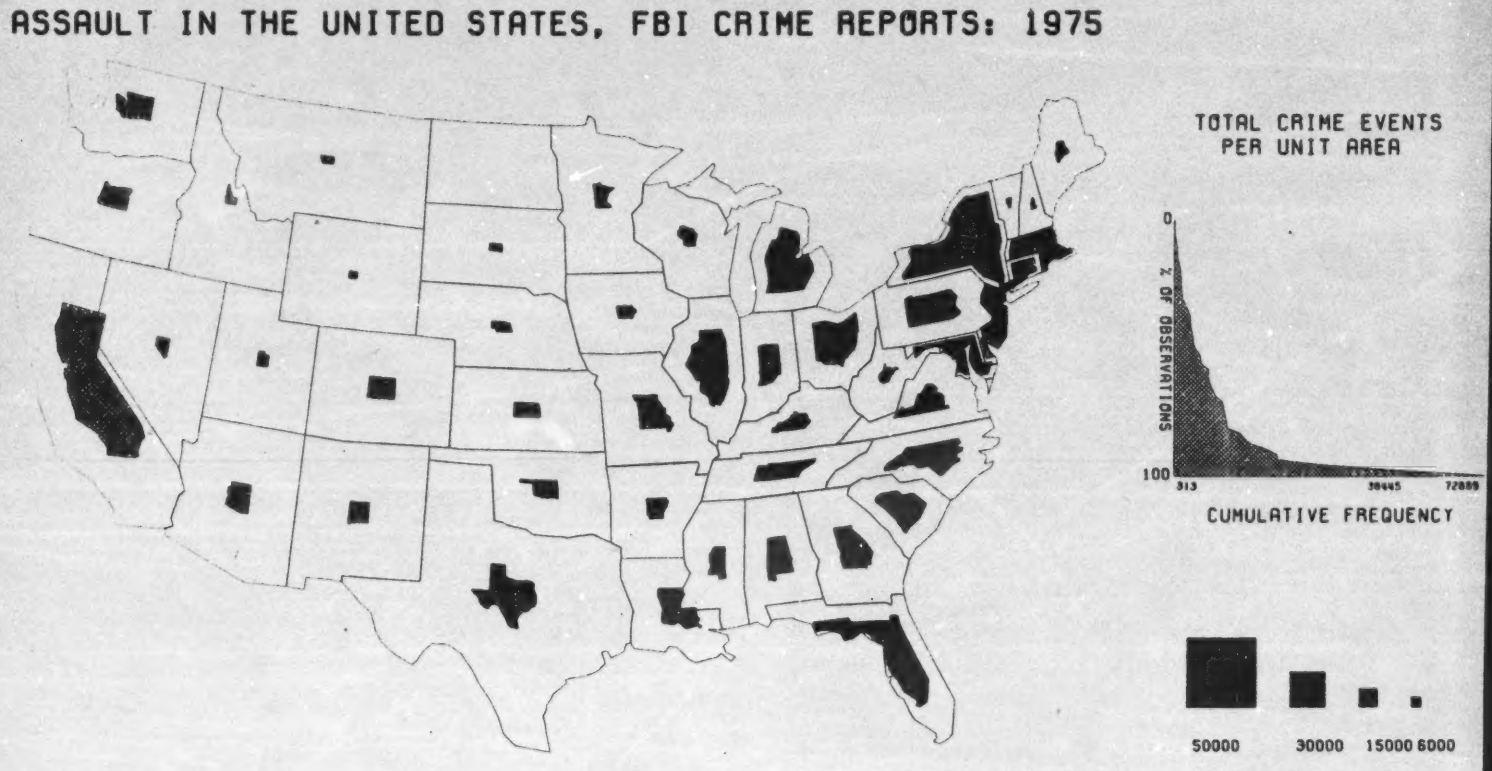

ASSAULT IN THE UNITED STATES, FBI CRIME REPORTS: 1975
TOTAL CRIME EVENTS
PER UNIT AREA
0
% OF OBSERVATIONS
100
313
38445
72889
CUMULATIVE FREQUENCY
50000
30000
15000
6000

Figure 11

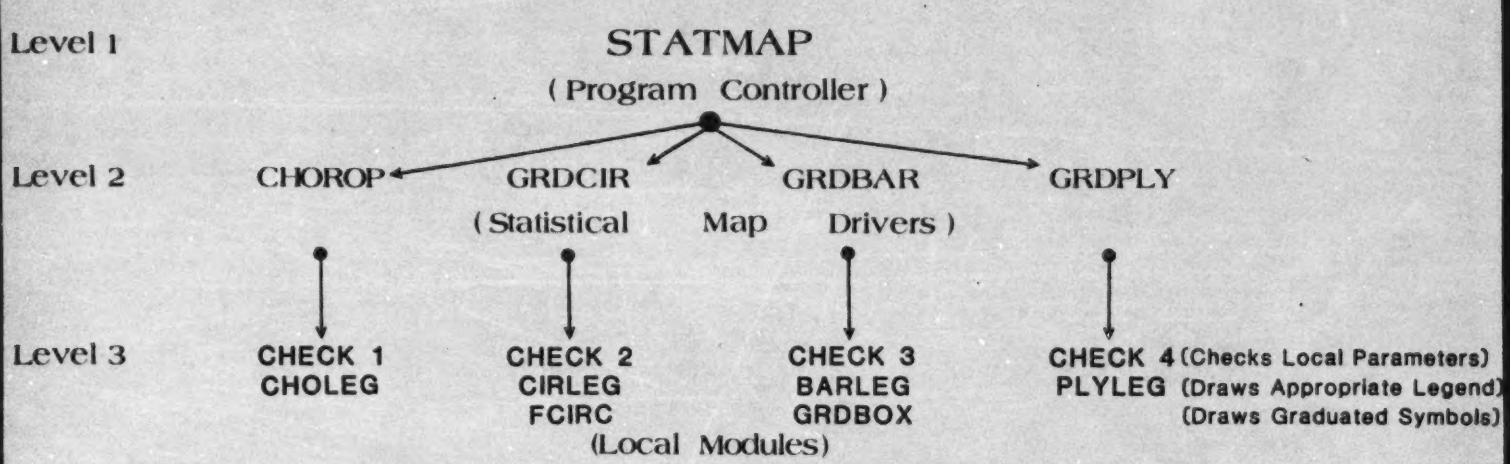
ANSI FORTRAN (international standards). The original program was designed and implemented at the Geographic Information Systems, State University of New York at Buffalo (Brassel, 1979; Marble, 1979). The extended version was developed at the laboratory for Cartographic and Spatial Analysis, The University of Akron as part of an overall project to provide mapping software for digitizing endeavors (Utano, 1979).

The program was designed and written in a modular format where specialized routines were written for both local and global tasks. In general, four major processing levels were built into the program. Figure 12 provides a schematic layout of this level hierarchy.

The first level contains the overall program controller. This controller is comparable to the main line driver of a FORTRAN program containing a set of subroutines. This routine processes the map type specification directive and routs control to the desired mapping subroutine. Processing control reverts to the program controller upon completion of a particular statistical map and the same procedure is resumed if another map is to be generated.

The second level contains the four statistical map processing drivers. Although each driver produces a different kind of graphic, they share in common many logical processing steps. In general, these four mapping modules undertake the following tasks:

(A) Initialize default parameters;
(B) Read global parameters;
(C) Read, echo, and check user-supplied local parameters
(D) Read in the X-Y coordinate and data file;
(E) Determine MIN/MAX data and X-Y values;
(F) Calculate appropriate symbolism for the areal units;
(G) For each areal unit:
    — read in the X-Y coordinate boundary
    — scale the coordinates to plotter space
    — draw the appropriate areal symbolism



Level 4

BOX	(draws neat line)	NDIGF	(finds number of digits to plot)
BREAK	(prints graph)	NUMBR	(special lettering routine)
CGRAV	(finds X-Y centers)	PARWR	(provides printer output)
CHECK	(checks global para.)	POLY	(draws non-valued polygrams)
DISTF	(finds shading density)	RDIN	(reads data files)
DMAXMN	(finds MIN-MAX)	RDPOLY	(reads locational data)
ERROR	(error processing)	ROUND	(rounds legend values)
FCURVE	(draws freq. graph)	SCALE	(fits X-Y values to plot)
FMTWR	(prints file formats)	SHADE	(shades polygons)
INPAR	(global input section)	SORT	(sorts data values)
MAXMN	(value MIN-MAX)	STITLE	(draws subtitles)
MINMAX	(X-Y MIN-MAX)	SYMBL	(special lettering routine)
MTITLE	(draws main title)	XYNEW	(finds scaled X-Y values)

(Global Modules)

Figure 12

(H) Plot additional 'non-valued' polygons if requested;
(I) Draw the legend, the cumulative frequency curve, the main and sub-titles, and the neat line;
(J) Set "MAP GENERATED" flag, increment the number of maps plotted during the present runstream, and write message that a map will be made on the plotter;
(K) Rout processing control to the program controller.

Level three contains processing routines designed for specific local tasks for each of the four mapping modules. Each map driver has from two to three of these local routines. The first set of routines is invoked to check the user-supplied local parameters. The second set of routines draws the appropriate legend symbolism. The last set is pertinent to the graduated-circle and -bar programs and they draw the graduated-symbols on the map body. The symbolism routines for the other two statistical maps is incorporated within the set of global processing subroutines since they are used by more than one mapping module.

The last level contains twenty-six global processing routines that are invoked by at least one of the four mapping modules. Figure 12 shows the processing mnemonics together with a brief description of their particular tasks.

Conclusion

Statistical mapping affords the researcher with a means of providing a visual replica of a statistical distribution. The representation medium is advantageous over the presentation of a spatial distribution in a verbal or tabular format for it furnishes the reader with a quick impression of any inherent spatial properties. To this end, production of statistical maps within a computerized framework has tended to enhance this information-communication advantage. Program STATMAP represents one example of how a computer can generate a number of different types of statistical maps in a relatively quick and inexpensive manner. The program is certainly not a panacea to some of the theoretical problems of statistical maps; however, it does demonstrate how a computer can be used to present spatial information in an optimal manner.